AI, Data Science & Health

Elaine Nsoesie, PhD
Assistant Professor of Global Health
Email: onelaine@bu.edu / Twitter: @ensoesie

BOSTON UNIVERSITY

How will you use AI and data science to improve health in Africa?

Health in Sub-Saharan Africa

- Life expectancy has increased from 53 years in 2000 to 64 years in 2017
    - Men is 62, Women is 66
- The number of children dying before the age five has decreased from 45% of all deaths in 1950 to 10% in 2017
- Overall decrease in mortality

Health in Sub-Saharan Africa

Leading causes of death for adults (15-49 years)

- AIDS
- Tuberculosis
- Malaria
- Maternal disorders
- Road injuries

Health in Sub-Saharan Africa

Impending epidemics of non-communicable diseases

- High blood pressure
- High blood sugar
- Obesity
- Alcohol use
- Diabetes
- Heart attacks
- Stroke

Exercise

Write down one health problem that you can solve with AI or data science. (10 mins)

For those of you interested, the goal is to

1. Come up with a simple/complicated problem that we can solve with AI
2. Think about what data we would need to solve this problem
3. Think about what methods we can use to solve this problem and why these methods are appropriate
4. Come up with a research plan
5. If you want to pursue that research plan beyond this course, I would be happy to advise you

Outline for the remainder of this session

Data

Methods

Ethics

Data

Types of digital data & uses

Mobile Apps & Crowdsourcing

Search

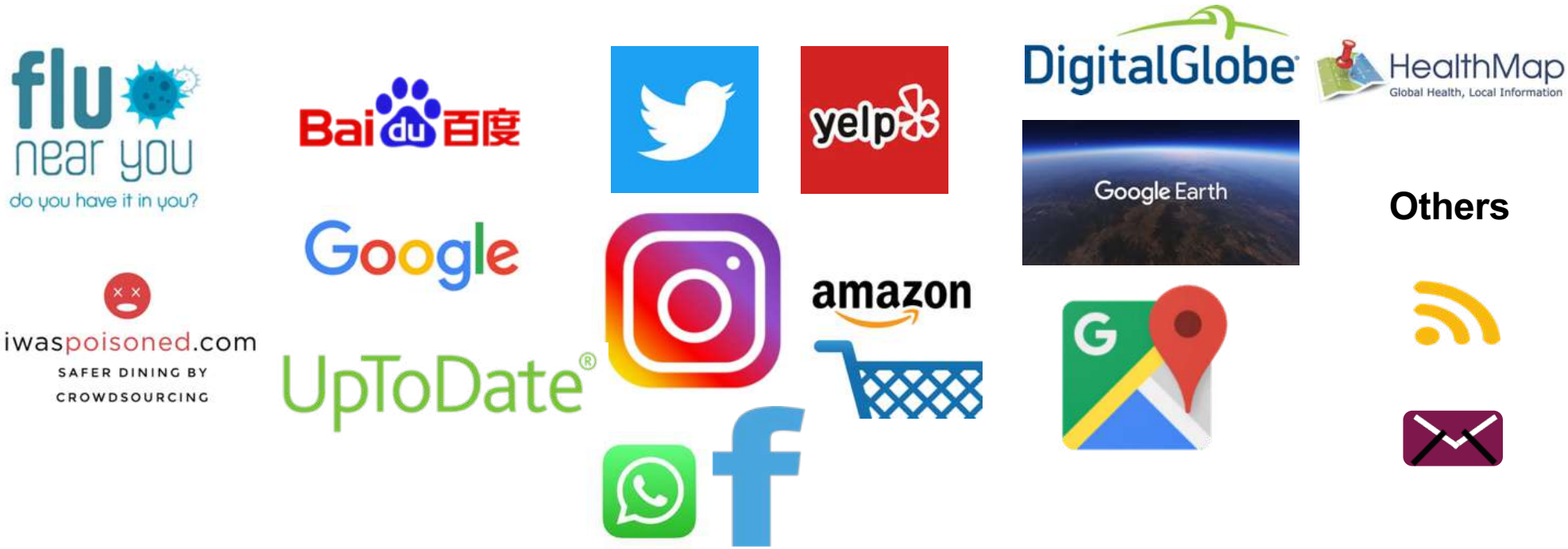
Social media

Consumer reviews

Remote sensing/place

News

Others



Some links to digital data sources

- trends.google.com
- https://dataforgood.fb.com/
- developer.twitter.com
- https://developers.google.com/maps

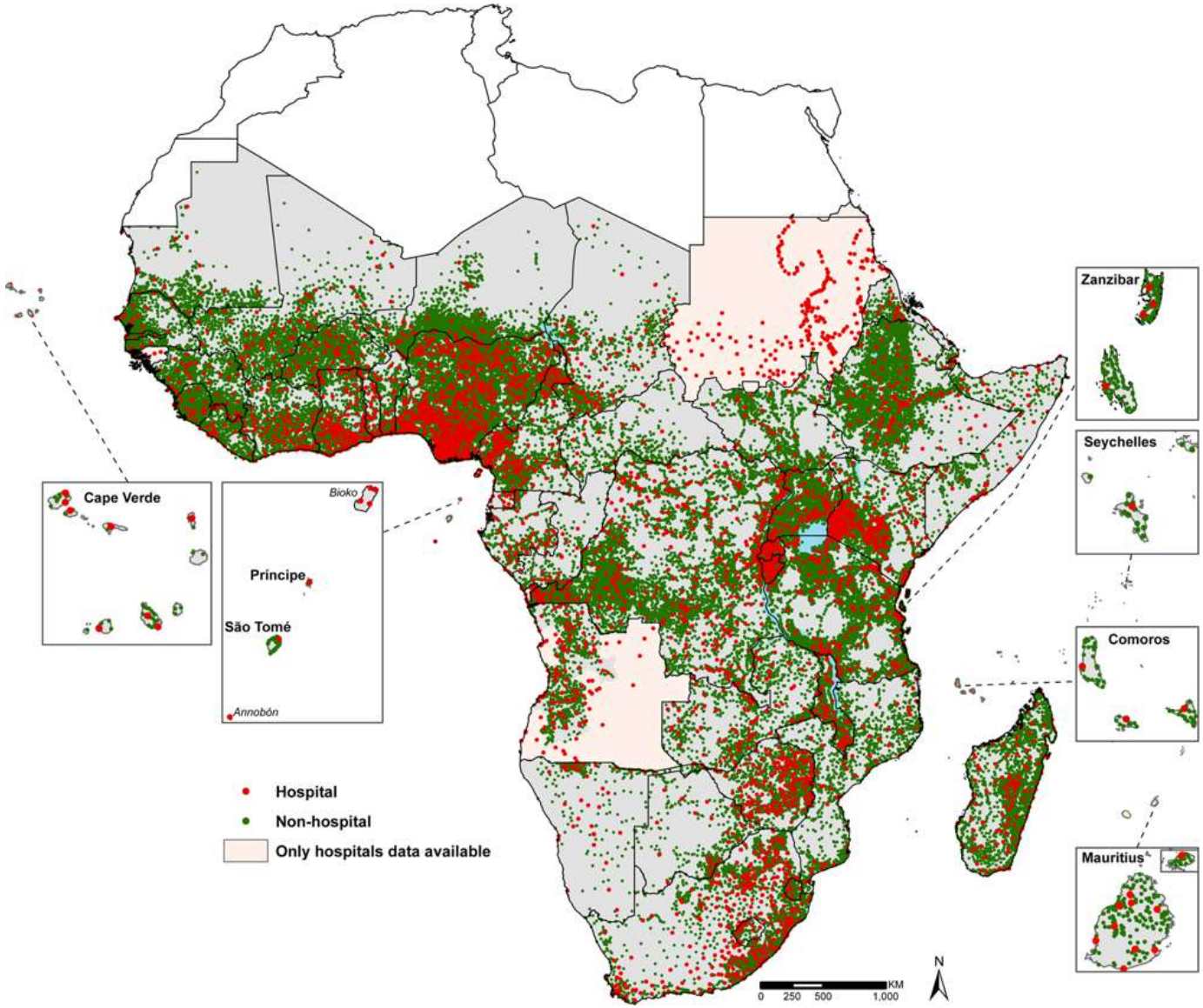
Surveys

- Country health surveys (e.g., family health survey)
- Census
- UNICEF
- DHS
- World Bank
- NGOs
- Published research studies

A spatial database of health facilities managed by the public health sector in sub Saharan Africa

Maina et al. (2019)

The distribution of 96,395 geocoded public health facilities in sub Saharan Africa. Red dots represent hospitals (n = 4,930) and green dots represent non-hospitals (n = 91,465). Facilities have been mapped on Global Administrative Unit Layers (GAUL) 2008 admin0 boundaries (http://www.fao.org/geonetwork/srv/en/main.home).



https://dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm 120%

The DHS Program
Demographic and Health Surveys

SEARCH | LOGIN | SELECT LANGUAGE

USAID FROM THE AMERICAN PEOPLE

WHO WE ARE | WHAT WE DO | WHERE WE WORK | DATA | PUBLICATIONS | TOPICS

The DHS Program > What-We-Do > Survey Types > Demographic and Health Survey (DHS)

What We Do

SURVEY TYPES

- Demographic & Health Survey (DHS)
- AIDS Indicator Survey (AIS)
- Service Provision Assessment (SPA)
- Malaria Indicator Survey (MIS)
- Key Indicators Survey (KIS)
- Other Quantitative Surveys
- Qualitative Research

DHS Overview

Demographic and Health Surveys (DHS) are nationally-representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health, and nutrition.

DHS Survey Types

There are two main types of DHS Surveys:

- **Standard DHS Surveys** have large sample sizes (usually between 5,000 and 30,000 households) and typically are conducted about every 5 years, to allow comparisons over time.
- **Interim DHS Surveys** focus on the collection of information on key performance monitoring indicators but may not include data for all impact evaluation measures (such as mortality rates). These surveys are conducted between rounds of DHS surveys and have shorter questionnaires than DHS surveys. Although nationally representative, these surveys generally have smaller samples than DHS surveys.

DHS Survey Topics

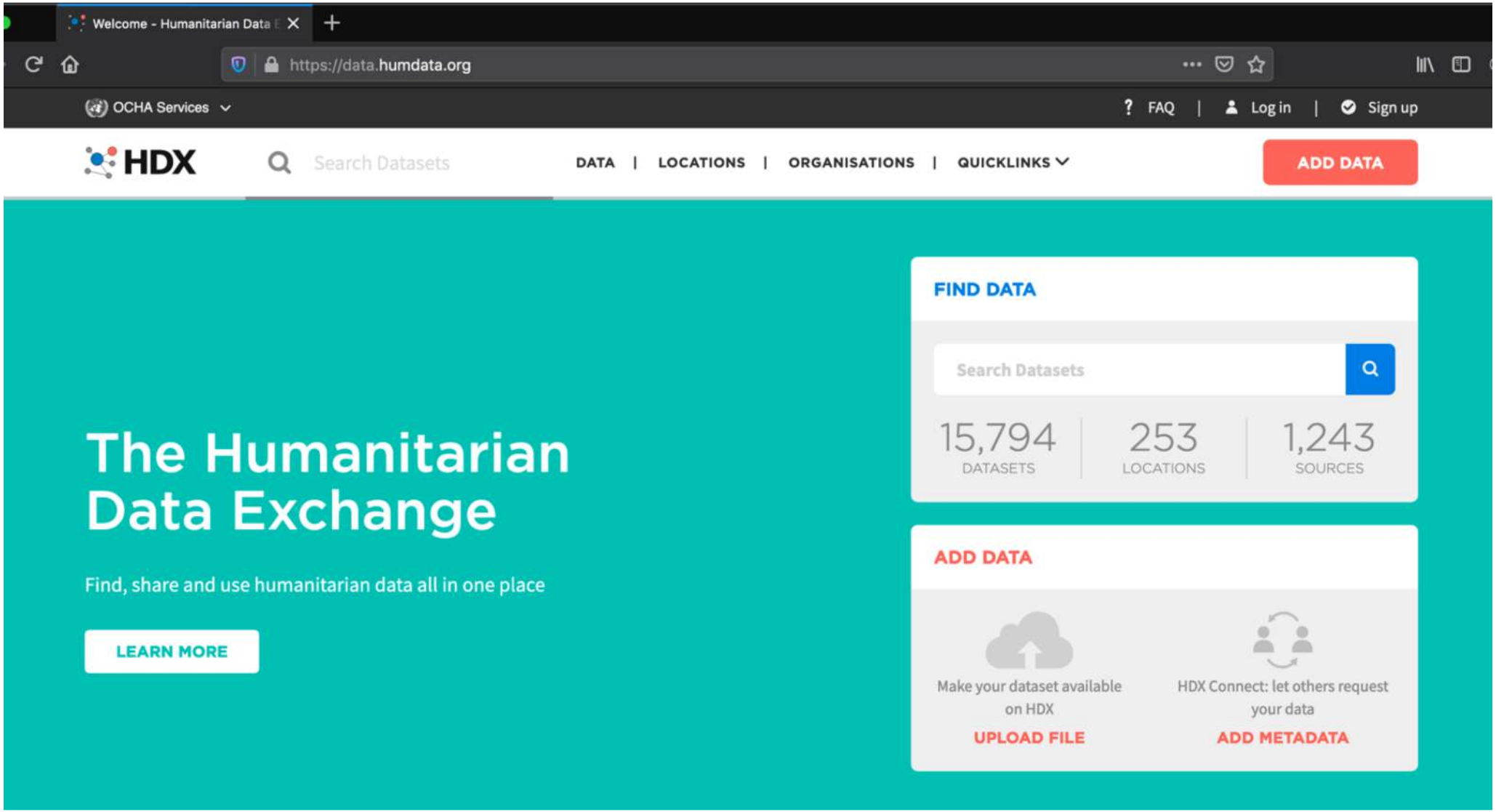
DHS Resources

- OVERVIEW
- METHODOLOGY
- QUESTIONNAIRES
- MANUALS

DHS Survey Topics

Information is available for the following topics, among others:

- Anemia - prevalence of anemia, iron supplementation
- Child Health - vaccinations, childhood illness, newborn care
- Domestic Violence (module) - prevalence of domestic violence and consequences of violence
- Education - literacy, attendance, highest level achieved
- Environmental Health - water, sanitation, cooking fuel
- Family Planning - knowledge and use of contraceptives
- Female Genital Cutting (module) - prevalence of and attitudes about female genital cutting
- Fertility and Fertility Preferences - total fertility rate, desired family size, marriage and sexual activity
- Gender/Domestic Violence - history of domestic violence, frequency and consequences of violence
- HIV/AIDS Knowledge, Attitudes, and Behavior - knowledge of HIV prevention, misconceptions, stigma, higher-risk sexual behavior, previous HIV testing
- HIV Prevalence - Prevalence of HIV by demographic and behavioral characteristics
- Household and Respondent Characteristics - electricity, housing quality, possessions, education and school attendance, age, sex, employment
- Infant and Child Mortality - infant and child mortality rates
- Malaria - ownership and use of mosquito nets, prevalence and treatment of fever, indoor residual spraying for mosquitoes
- Maternal Health - antenatal, delivery and postnatal care
- Maternal Mortality (module)- maternal mortality ratio
- Nutrition - child feeding practices, vitamin supplementation, anthropometry, anemia, salt iodization
- Tobacco Use - tobacco use, exposure to second-hand smoke
- Unmet Need for family planning
- Wealth - division of households into 5 wealth quintiles to show relationship between wealth, population and health indicators
- Women's Empowerment - gender attitudes, women's decision making power, education and employment of men vs. women


Welcome - Humanitarian Data E
https://data.humdata.org
OCHA Services
FAQ
Log in
Sign up
HDX
Search Datasets
DATA
LOCATIONS
ORGANISATIONS
QUICKLINKS
ADD DATA
The Humanitarian Data Exchange
Find, share and use humanitarian data all in one place
LEARN MORE
FIND DATA
Search Datasets
15,794
DATASETS
253
LOCATIONS
1,243
SOURCES
ADD DATA
Make your dataset available on HDX
UPLOAD FILE
HDX Connect: let others request your data
ADD METADATA

Data | DataBank

https://databank.worldbank.org/databases.aspx

Help us improve this section of the site. Can we get your feedback? Click here

THE WORLD BANK

Home About Data Research Learning News Projects & Operations Publications Countries Topics English

This page is in English Español Français عربي 中文

DataBank

Log in Now TWEETS LIKE SHARE +

- DataBank Home
- Databases
- Create Report
- Saved Reports
- Metadata Glossary

WHAT'S NEW

Joint External Debt Hub was updated on November 20, 2019

Doing Business was updated on November 20, 2019

Worldwide Governance Indicators was updated on November 7, 2019

World Development Indicators was updated on October 28, 2019

Explore databases

Type keywords to filter database names

Filter by Topic

Sort by Most Used | Alphabetical | Last Updated

Showing results 10 of 78 Database preview ON OFF

World Development Indicators Public

World Development Indicators (WDI) is the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global See more +

External Debt and Financial Flows statistics, Heath statistics, Gender, Economy, Social Data

Last Updated:10/28/2019

Statistical Capacity Indicators Public

Statistical Capacity Indicators provides information on various aspects of national statistical systems of developing countries, including an overall country-level statistical capacity indicator.

Last Updated:01/24/2019

https://africaopendata.org/dataset?q=&sort=views_recent+desc

openAFRICA
by Code for Africa

Datasets Organisations Groups About Data Requests Log in

Search datasets...

5,513 datasets found

Order by: Popular

Uganda Districts Shape Files
Updated August 19, 2015 | Created May 23, 2013

Uganda Districts Shape Files

Census 2011 - Shape Files
Updated August 17, 2015 | Created November 23, 2012

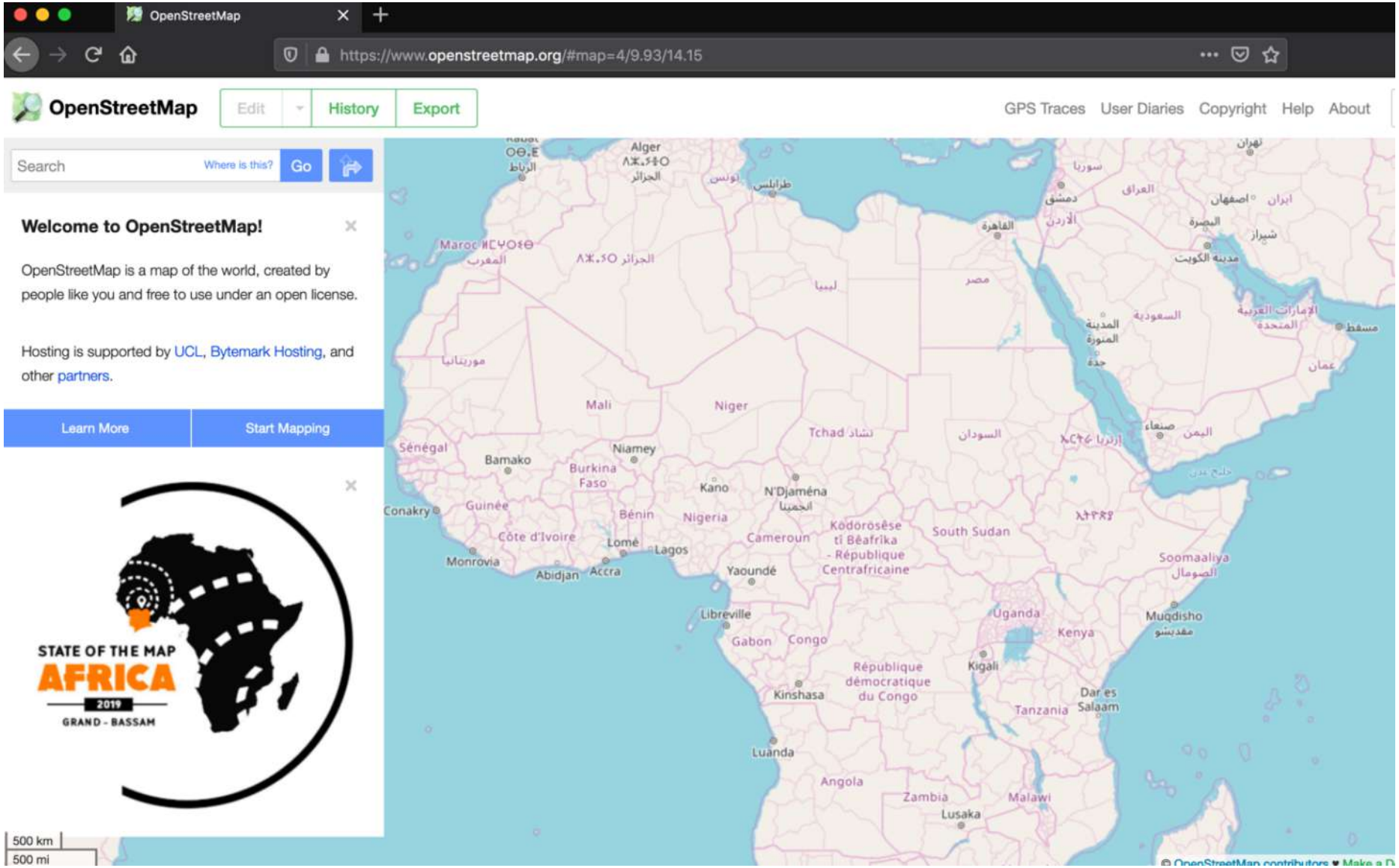
This dataset has no description

Trust Fund Grants Committed & Disbursed - Eritrea
Updated August 19, 2015 | Created November 4, 2012

A Recipient-executed Grant is a Trust Fund Grant that is provided to a third party under a grant agreement, and for which the Bank plays an operational role - i.e., the Bank...

World Bank Contract Award 2010-2012 - Niger
Updated August 19, 2015 | Created November 4, 2012

This set of contract awards includes data on commitments against contracts that were reviewed by the Bank before they were awarded (prior-reviewed


OpenStreetMap
https://www.openstreetmap.org/#map=4/9.93/14.15
OpenStreetMap
Edit
History
Export
GPS Traces
User Diaries
Copyright
Help
About
Search
Where is this?
Go
Welcome to OpenStreetMap!
OpenStreetMap is a map of the world, created by people like you and free to use under an open license.
Hosting is supported by UCL, Bytemark Hosting, and other partners.
Learn More
Start Mapping
STATE OF THE MAP
AFRICA
2019
GRAND - BASSAM
Alger
Mali
Niger
Tchad
Sénégal
Bamako
Niamey
Burkina Faso
Kano
N'Djaména
Conakry
Guinée
Bénin
Nigeria
Côte d'Ivoire
Lomé
Lagos
Cameroun
Monrovia
Abidjan
Accra
Yaoundé
Libreville
Gabon
Congo
Kinshasa
Luanda
Angola
Zambia
Lusaka
Malawi
Tanzania
Dar es Salaam
Kigali
Uganda
Kenya
South Sudan
Soomaaliya
Muqdisho
Kodorosêse ti Bêafrîka - République Centrafricaine
République démocratique du Congo
Maroc
Le Caire
Khartoum
Addis Abeba
Asmara
Sana'a
Riyadh
Jeddah
Medina
Dubai
Muscat
Tehran
Baghdad
Damascus
Amman
Beirut
Tripoli
Tunis
Rabat
500 km
500 mi

https://www.naiis.ng

NAIIS 2018 ...getting to zero

Home | About NAIIS | People | Resource | Gallery

The 2018 Nigeria HIV/AIDS Indicator and Impact Survey (NAIIS)

is a cross-sectional survey that assessed the prevalence of key human immunodeficiency virus (HIV)-related health indicators.

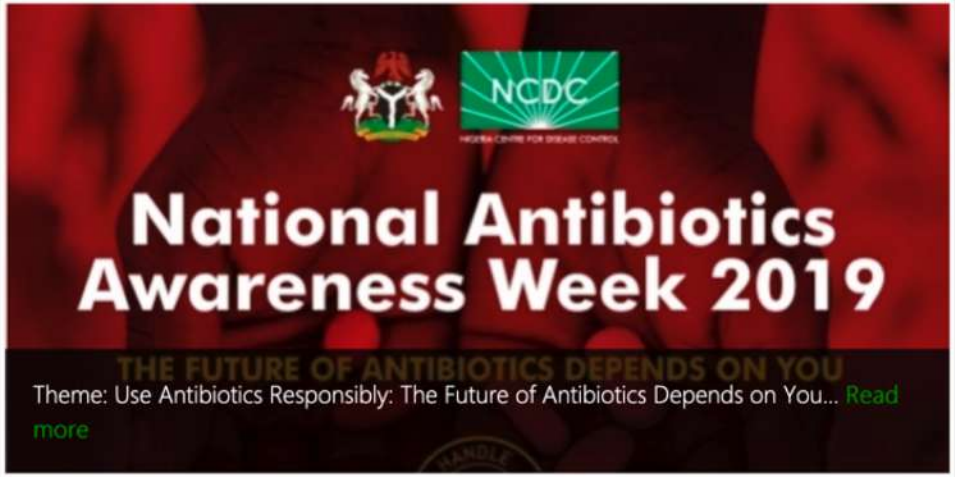
https://ncdc.gov.ng

NCDC
NIGERIA CENTRE FOR DISEASE CONTROL

Nigeria Centre for Disease Control

Protecting the health of Nigerians

Home | About | Publications | Diseases | News/Media | Training/Events | Projects | Jobs | Preparedness | Dashboard | Contact



- 2019 National Antibiotics Awareness Week
- Nigeria One Health Strategic Plan
- Public Health Advisory on Cholera: 5 Tips for Prevention
- Cholera Advisory for Healthcare Workers and Providers

Exercise

What type(s) of data and what sources would you use to solve the health problem you previously identified? (10 mins)

Methods

Satellite Images and deep learning

About 55% of the world's population lives in urban areas. This number is expected to increase to 68% by 2050.

United Nations

The field of urban health which started around the turn of the 21st century is focused on the study of how and why cities influence health.

The foods we have access to

The foods we eat

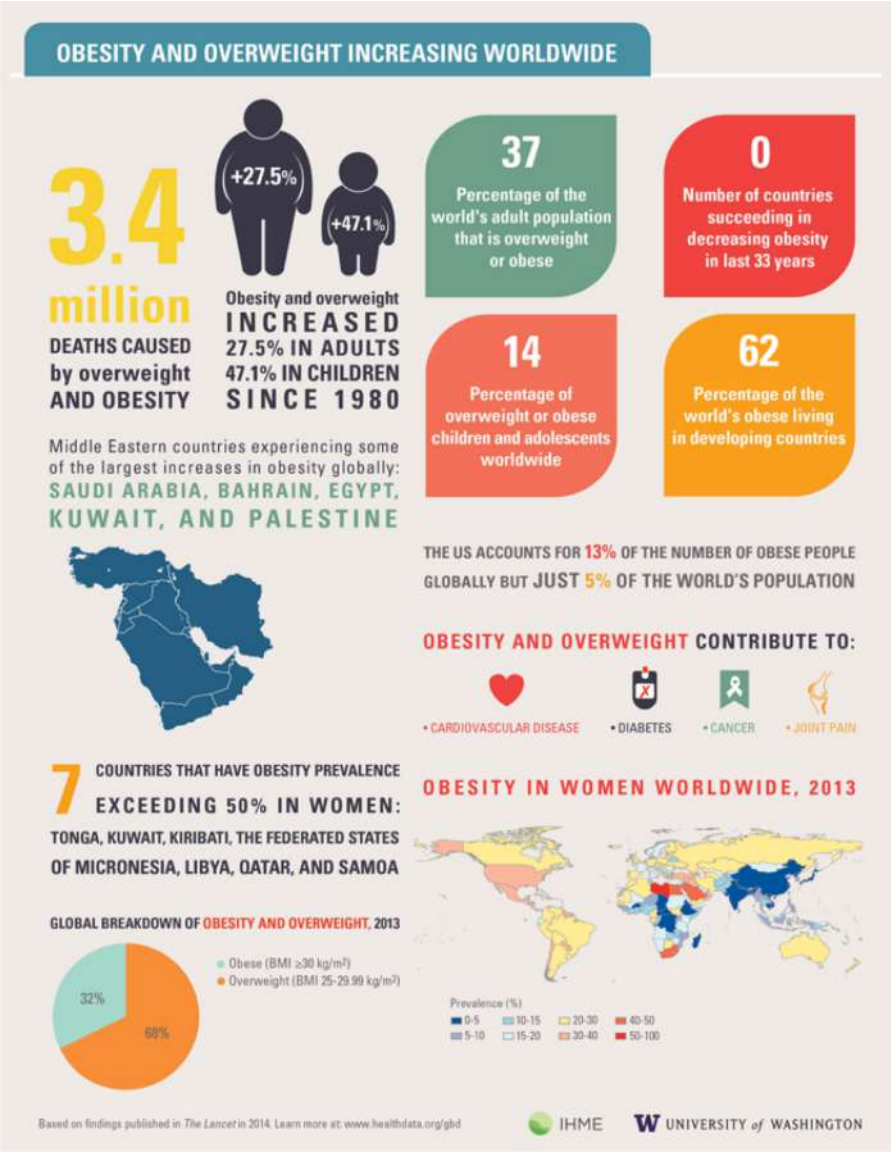
Images from unsplash.com

All neighborhoods are not created equal

Drivers of health in cities

The physical environment
The social environment
Access to health and social services

- Obesity is a complex health issue
- Multiple factors have been linked to obesity
- In the US, obesity affects about one-third of the adult population



Many environmental factors have been linked to obesity

But there are differences in measures and measurements making it difficult to compare findings across cities

Obesity prevalence across neighborhoods in six cities

- Memphis, Tennessee (6th)
- Seattle-Bellevue-Tacoma, Washington (32nd)
- San Antonio, Texas (8th)
- Los Angeles, CA (47th)

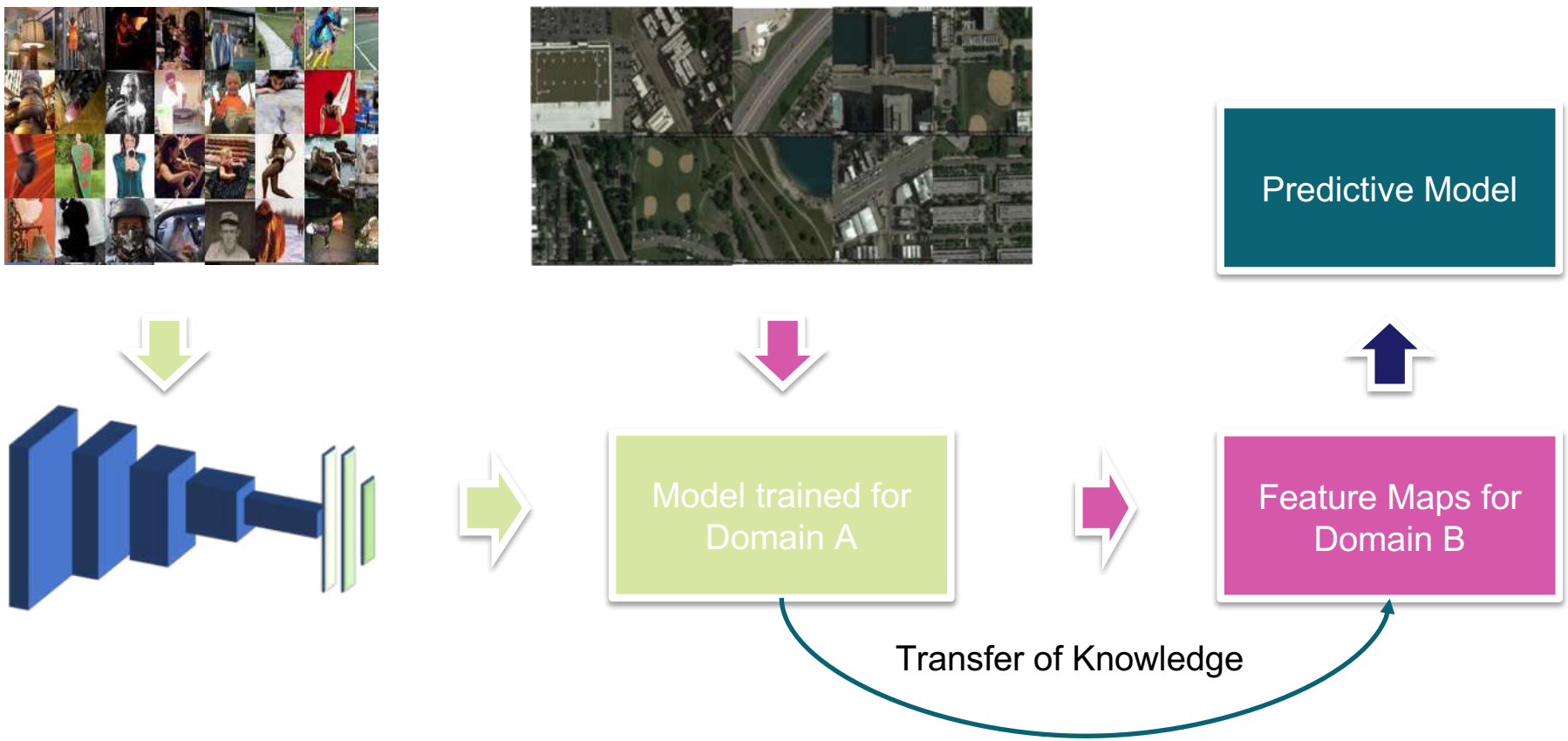
- Estimates of obesity prevalence from CDC
- Points of interest
- Per capita income
- Satellite images from Google

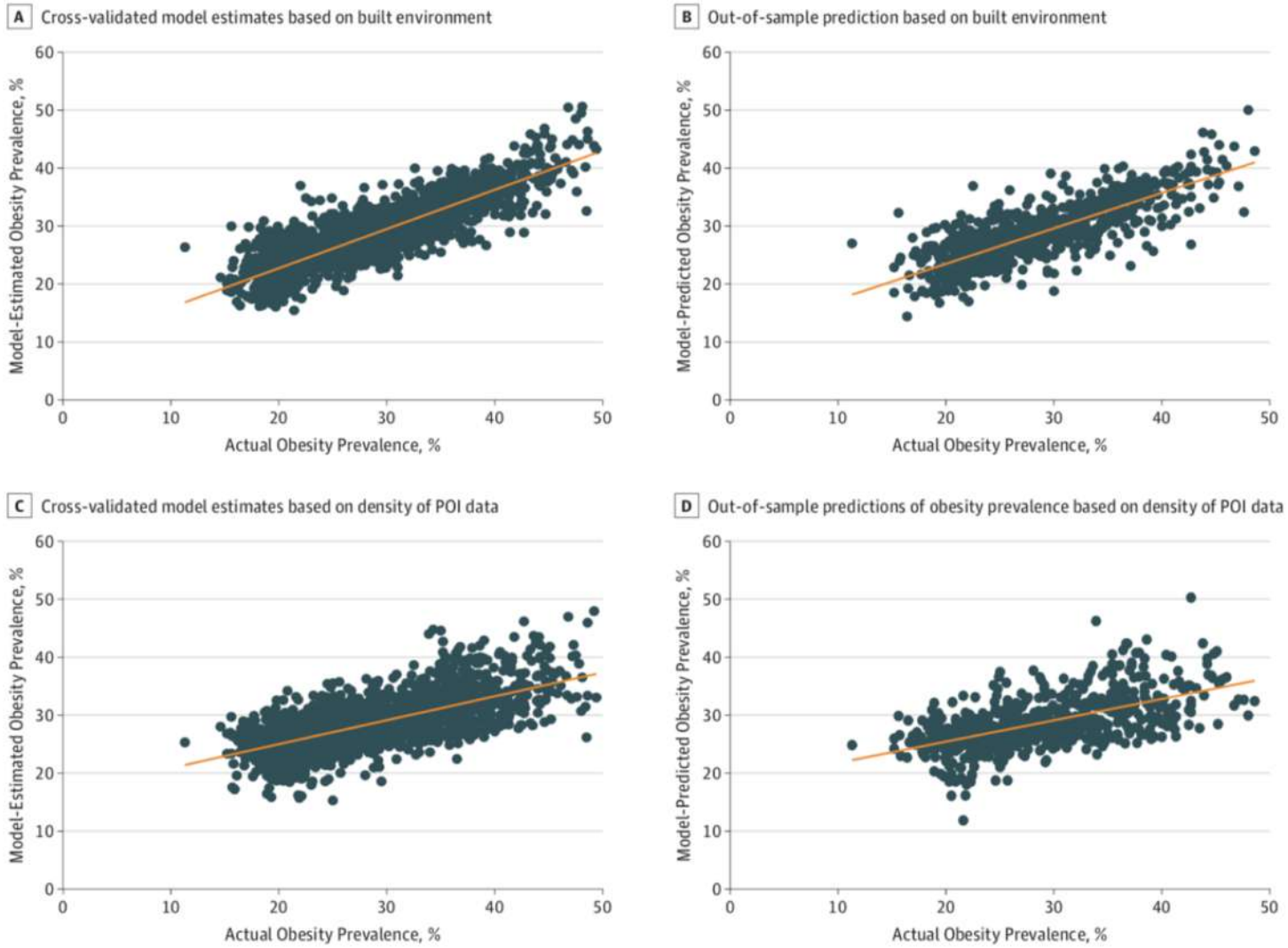
~150,000 recent satellite images were downloaded from Google Static Maps API

Zoom level 8

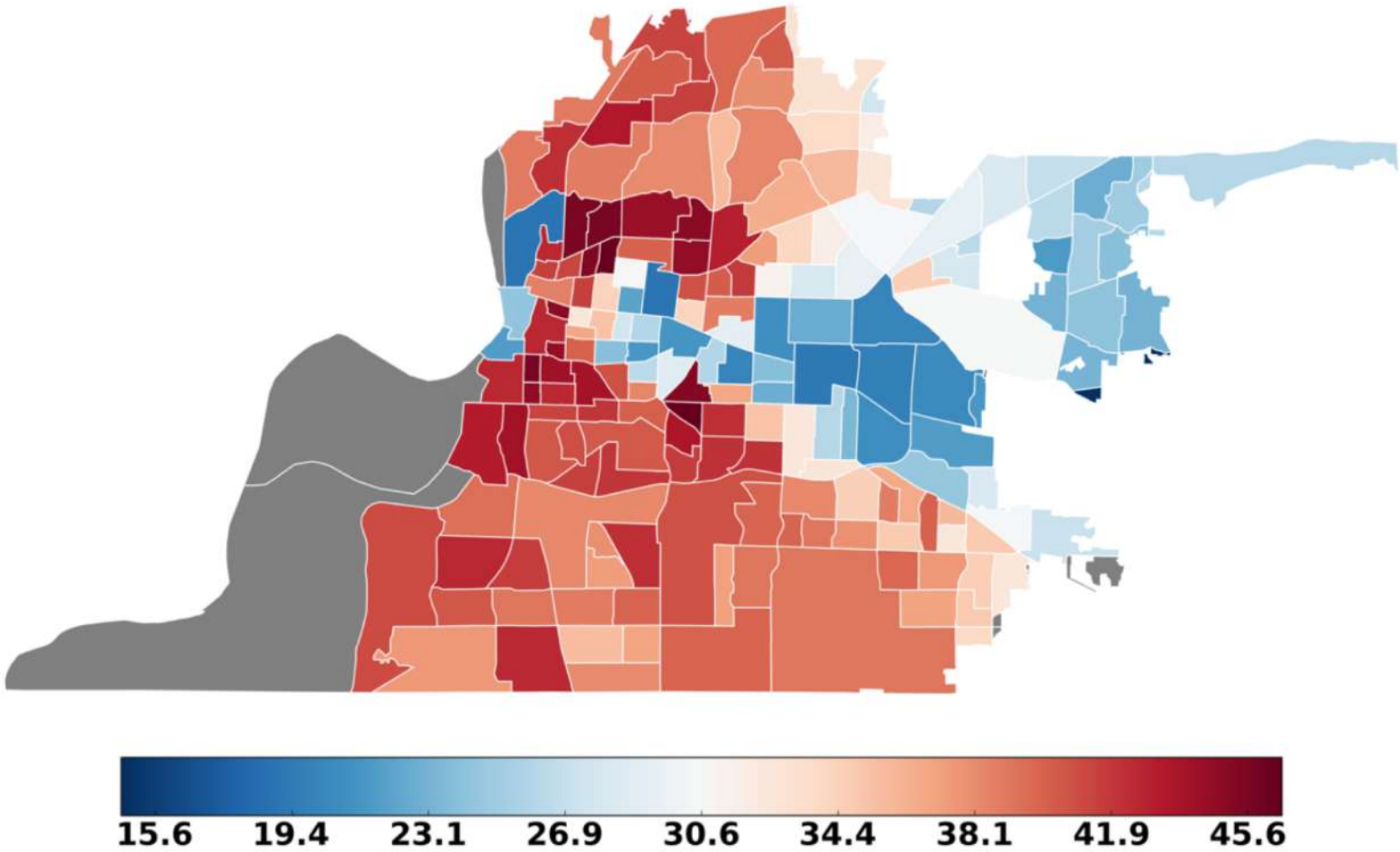
Data – Satellite Images

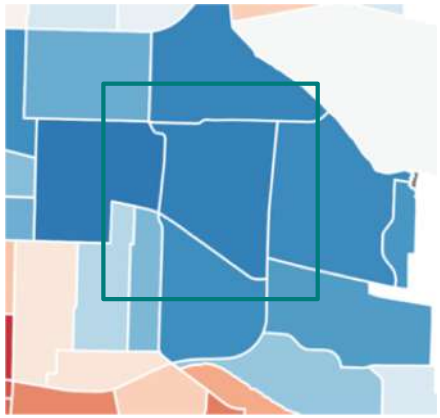
Transfer Learning with VGG-CNN-F





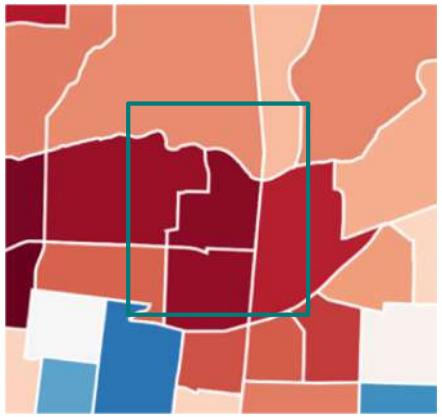
Built environment information is extracted from satellite images. POI indicates point of interest.


15.6
19.4
23.1
26.9
30.6
34.4
38.1
41.9
45.6



Google ©2017 , DigitalGlobe, State of Arkansas, USDA Farm Service Agency

Google ©2017 , DigitalGlobe, State of Arkansas, USDA Farm Service Agency

Google ©2017 , DigitalGlobe, State of Arkansas, USDA Farm Service Agency


Google ©2017 , DigitalGlobe, State of Arkansas, USDA Farm Service Agency

Google ©2017 , DigitalGlobe, State of Arkansas, USDA Farm Service Agency

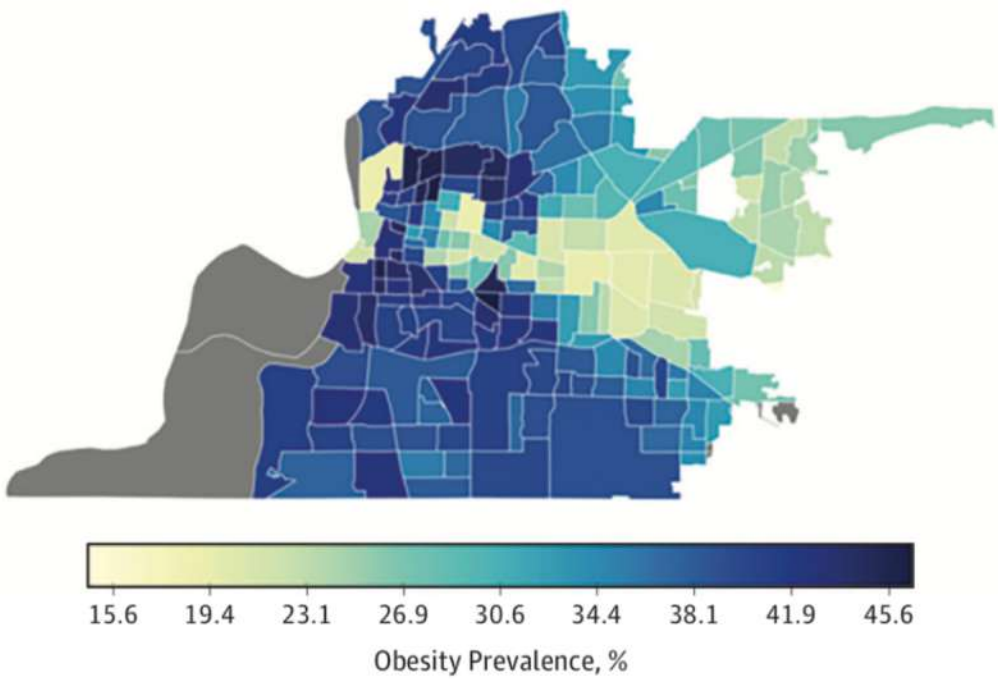
Google ©2017 , DigitalGlobe, State of Arkansas, USDA Farm Service Agency

San Antonio, Texas

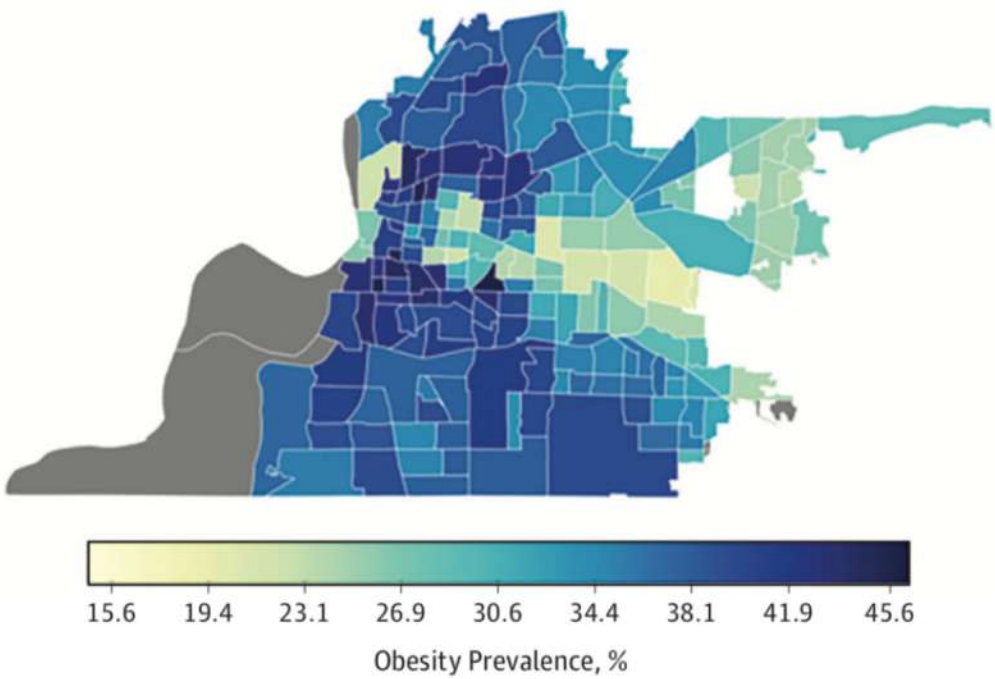
N=311

Actual

A Memphis, Tennessee



Estimate



Mapping

Chikungunya Disease Transmission and Implications for Surveillance

Chikungunya Disease Transmission and Implications for Surveillance

- Chikungunya virus was first identified in 1953 in Tanganyika (Tanzania)
- Chikungunya is an acute febrile illness
- Symptoms: incapacitating joint pain, high fever and skin rash
- Endemic in several Asian and African countries

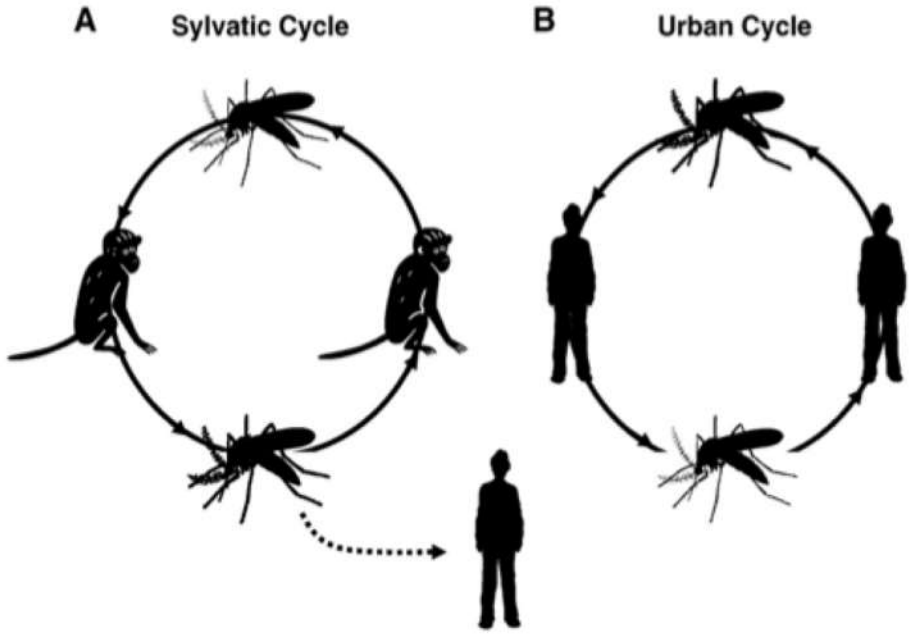
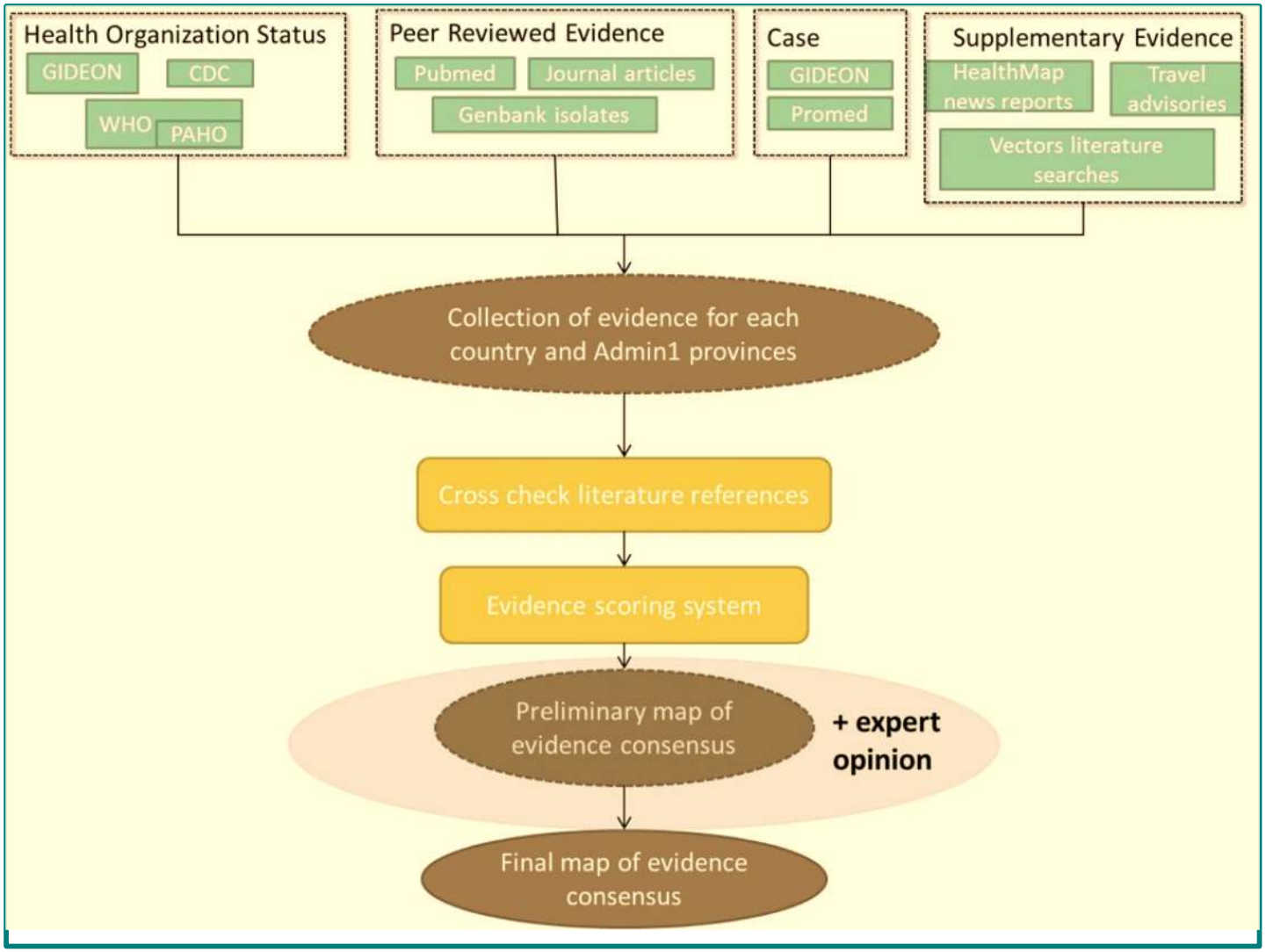


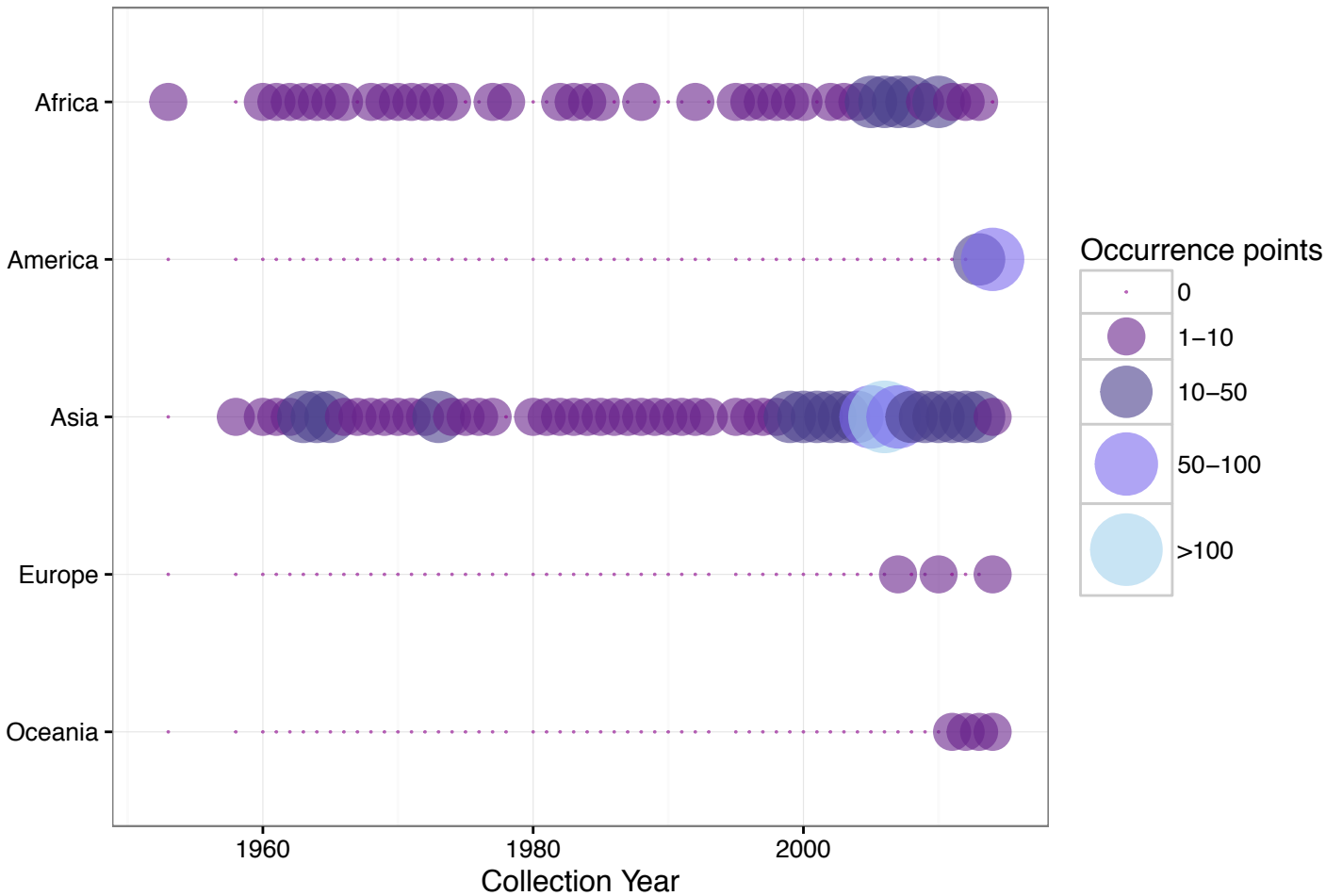
Fig 1. (A) CHIKV in Africa is maintained in a sylvatic cycle involving forest dwelling *Aedes* spp. mosquitoes and nonhuman primates. When sylvatic mosquito densities increase, often during periods of heavy rainfall, small human epidemics or sporadic human cases may occur. (B) In urban settings, CHIKV circulates in a man-mosquito-man cycle vectored principally by the anthropophilic *A aegypti* mosquito. Although *A albopictus* has been considered an accessory vector, some recent urban outbreaks have been vectored primarily by this mosquito.


Health Organization Status
GIDEON
CDC
WHO
PAHO
Peer Reviewed Evidence
Pubmed
Journal articles
Genbank isolates
Case
GIDEON
Promed
Supplementary Evidence
HealthMap news reports
Travel advisories
Vectors literature searches
Collection of evidence for each country and Admin1 provinces
Cross check literature references
Evidence scoring system
Preliminary map of evidence consensus
+ expert opinion
Final map of evidence consensus

Evidence Consensus

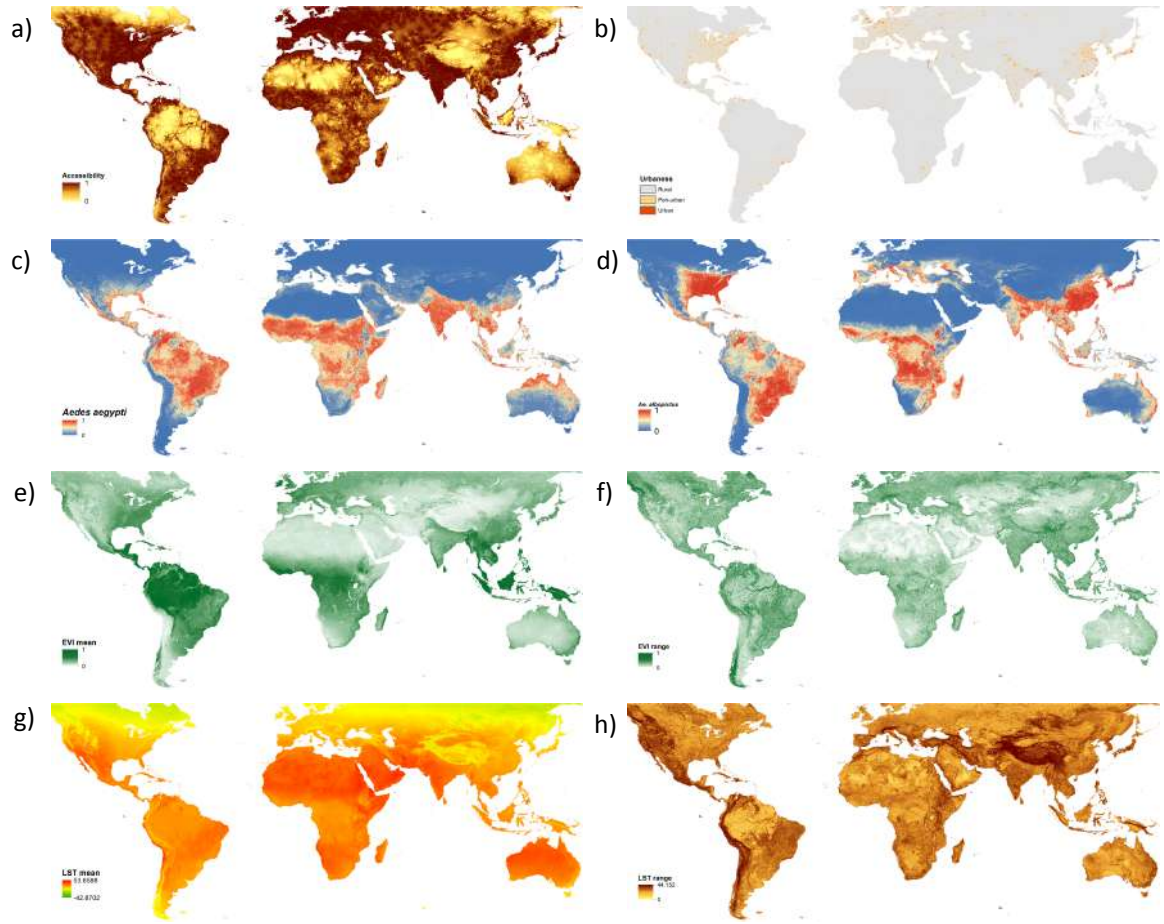
- Summary statistic representing confidence in presence or absence of Chikungunya for a given political region
- Evidence consensus score estimated at subnational level for France, Italy, India, Brazil, Mexico, Argentina, United States of America, and China
- Evidence consensus was mapped onto seven equidistant categories: complete presence/absence, good presence/absence, moderate presence/absence, and intermediate

Extraction of occurrence points



Extraction of covariates

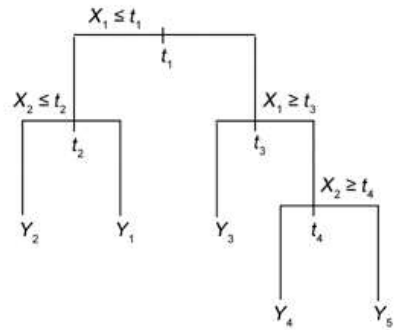
a) Urban accessibility
b) Urban, peri-urban and rural areas
c) *Aedes aegypti* suitability
d) Ae. *albopictus* suitability
e) EVI mean values
f) EVI range
g) LST mean values
h) LST range

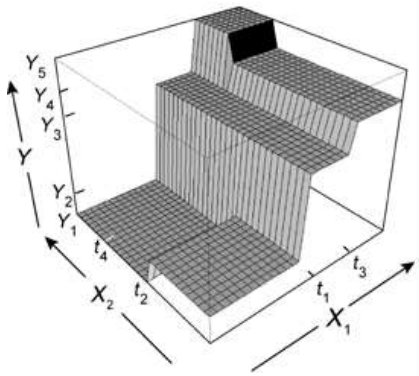


Model Specifications

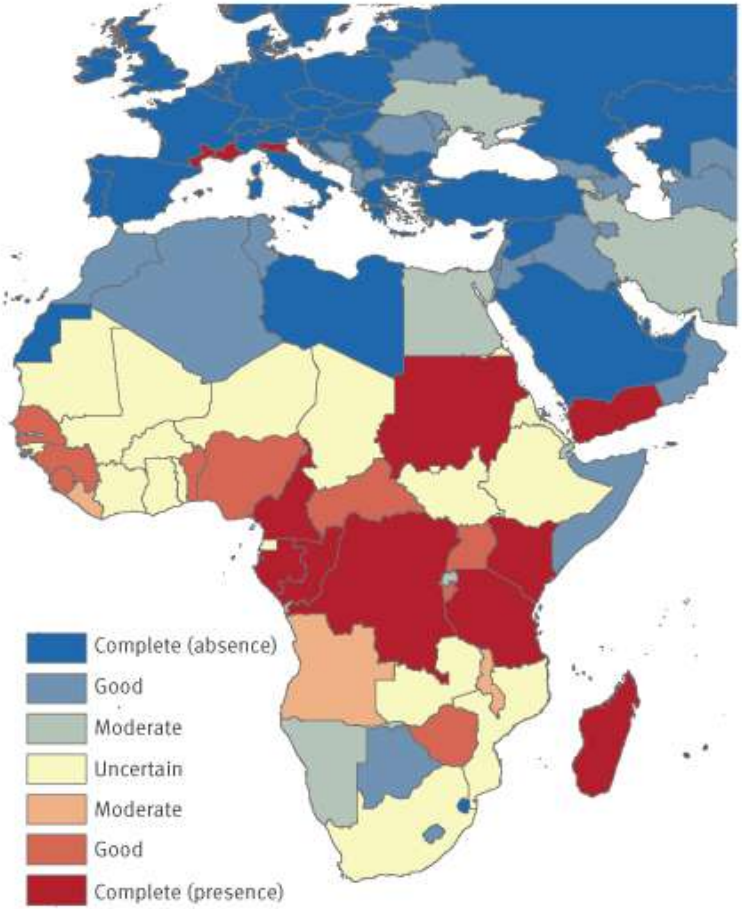
Boosted regression tree (BRT)

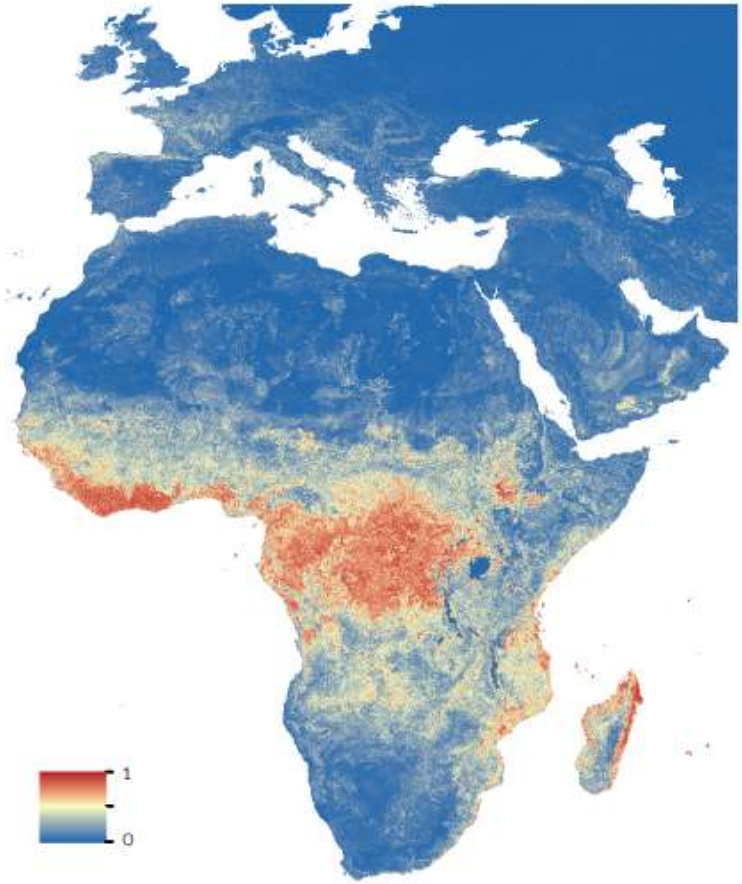
- Combining both, machine learning approaches and regression techniques
- Predictor variables can be of any type (numeric, binary, categorical, etc.)
- Boosting: a forward, stagewise procedure
- Outperforms MaxEnt, GARP, BIOCLIM

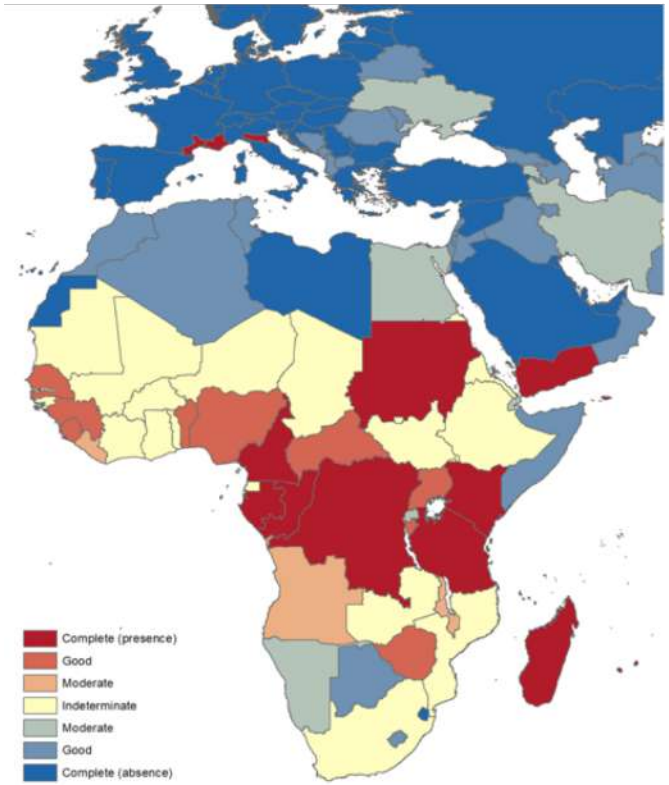




seeg
spatial ecology and epidemiology group


Complete (absence)
Good
Moderate
Uncertain
Moderate
Good
Complete (presence)


1
0



Africa & Europe

54% of African countries had good or complete evidence for CHIKV transmission

Insufficient data to determine presence or absence for 30% of countries in Africa

Local transmission of CHIKV in Europe has been reported in Ravenna, northern Italy in 2007; the southeastern French city of Fréjus in 2010 and Montpellier, southern France in 2014

Natural Language Processing

Open access Original article

BMJ Open Sport & Exercise Medicine

Social media captures demographic and regional physical activity

Nina Cesare,[1,2] Quynh C Nguyen,[3] Christan Grant,[4] Elaine O Nsoesie[1,2]

To cite: Cesare N, Nguyen QC, Grant C, *et al*. Social media captures demographic and regional physical activity. *BMJ Open Sport & Exercise Medicine* 2019;**5**:e000567. doi:10.1136/bmjsem-2019-000567

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136bmjsem-2019-000567).

Accepted 2 July 2019

ABSTRACT

Objectives We examined the use of data from social media for surveillance of physical activity prevalence in the USA.

Methods We obtained data from the social media site Twitter from April 2015 to March 2016. The data consisted of 1 382 284 geotagged physical activity tweets from 481 146 users (55.7% men and 44.3% women) in more than 2900 counties. We applied machine learning and statistical modelling to demonstrate sex and regional variations in preferred exercises, and assessed the association between reports of physical activity on Twitter and population-level inactivity prevalence from the US Centers for Disease Control and Prevention.

Results The association between physical inactivity tweet patterns and physical activity prevalence varied by sex and region. Walking was the most popular physical activity for both men and women across all regions

What are the findings?

- ► Men mentioned engaging in higher intensity physical activities than women, which agrees with previous studies suggesting that women are less likely to meet recommendations for aerobic physical activity.
- ► There were differences in the types of physical activities reported across the four US regions.

How might it impact clinical practice?

- ► Differences in the types of physical activities reported across sex and regions in the US can encourage discussions between clinicians and patients regarding exercise choices for weight loss and cardiovascular health.

Demographics

208.9 Million

Social media users in the US (2017)

91%
Of 18-29 year olds use **Youtube**

78%
Of 30-49 year olds use **Facebook**

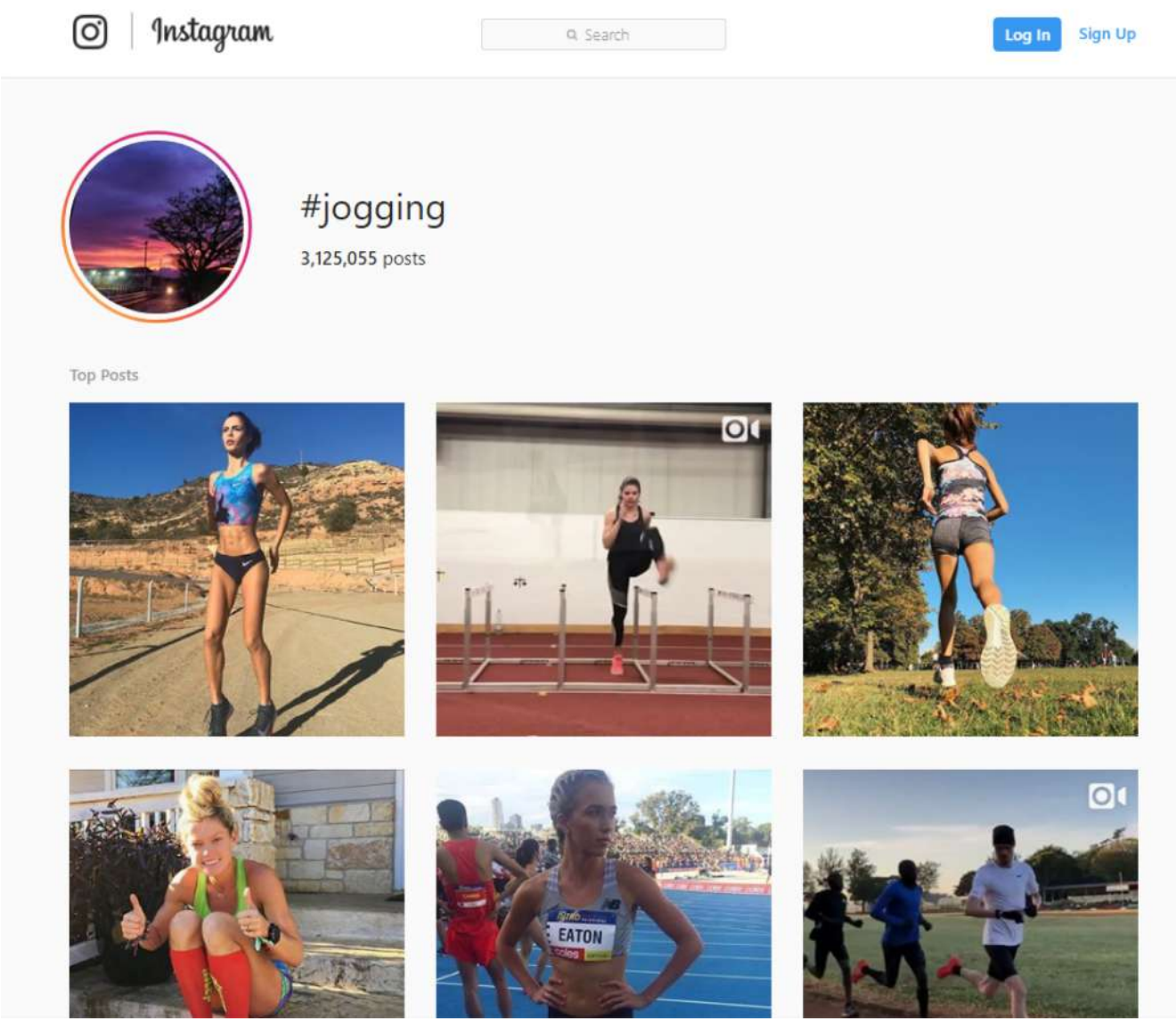
20%
Of users making <$30k use **Twitter**

78%
Of Hispanics use **Youtube**

Pew Research Center "*Social Media Use in 2018*"

Social media provides a window into the built environment, interactions, access, likes/dislikes, lifestyle, behaviors, choices

Social media can provide timely and low-cost information on **community health**



Social media can provide data on

1. Activities

2. Sentiment

3. Scale

4. Changes

Does it matter that we do not have demographics from some social media platforms?

Understanding Demographic Disparities

- Important gender differences in risk factors
- Important age differences in how people frame and discuss risk factors
- Spatial representation – higher population, higher income counties typically have more data available

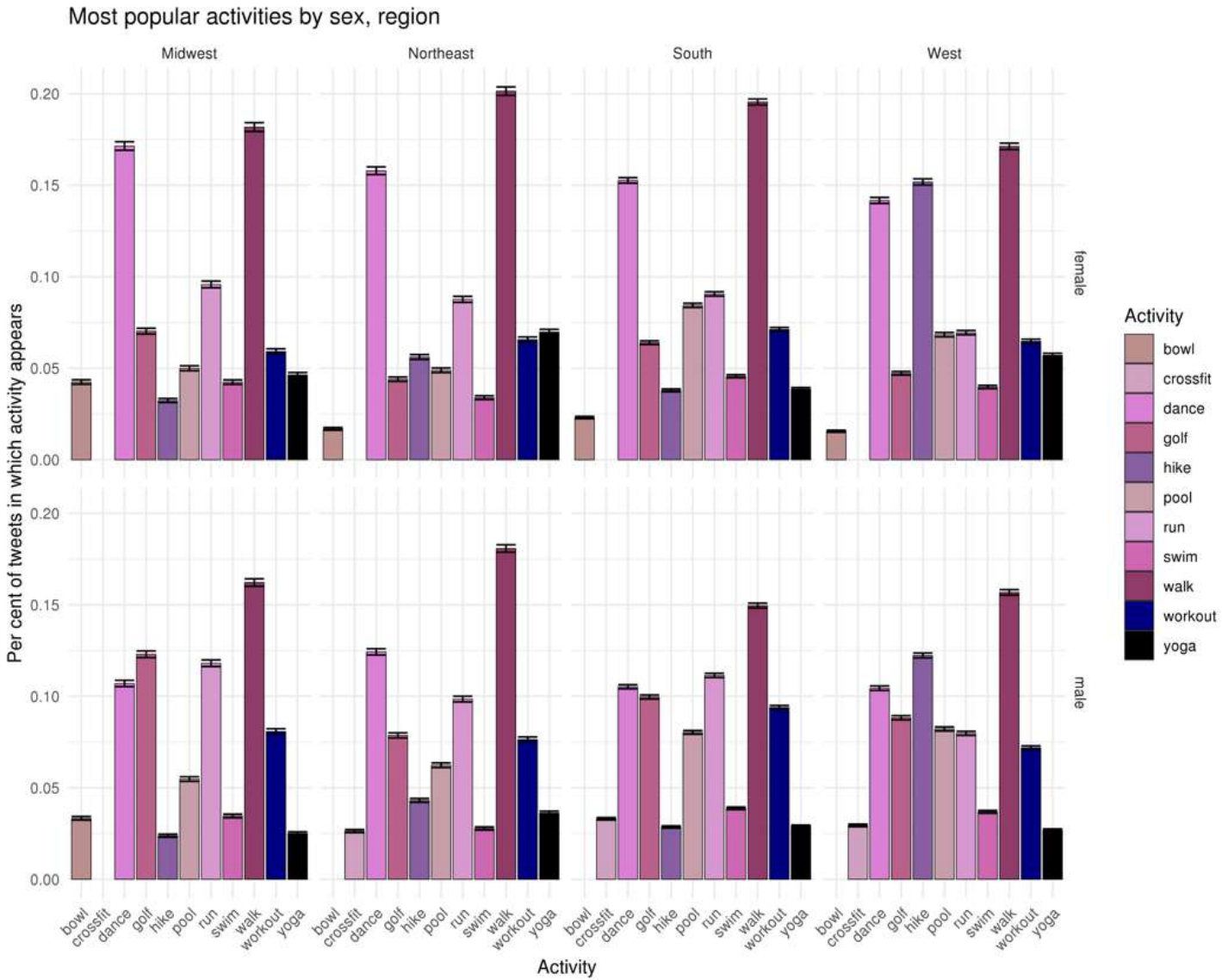
1,382,284

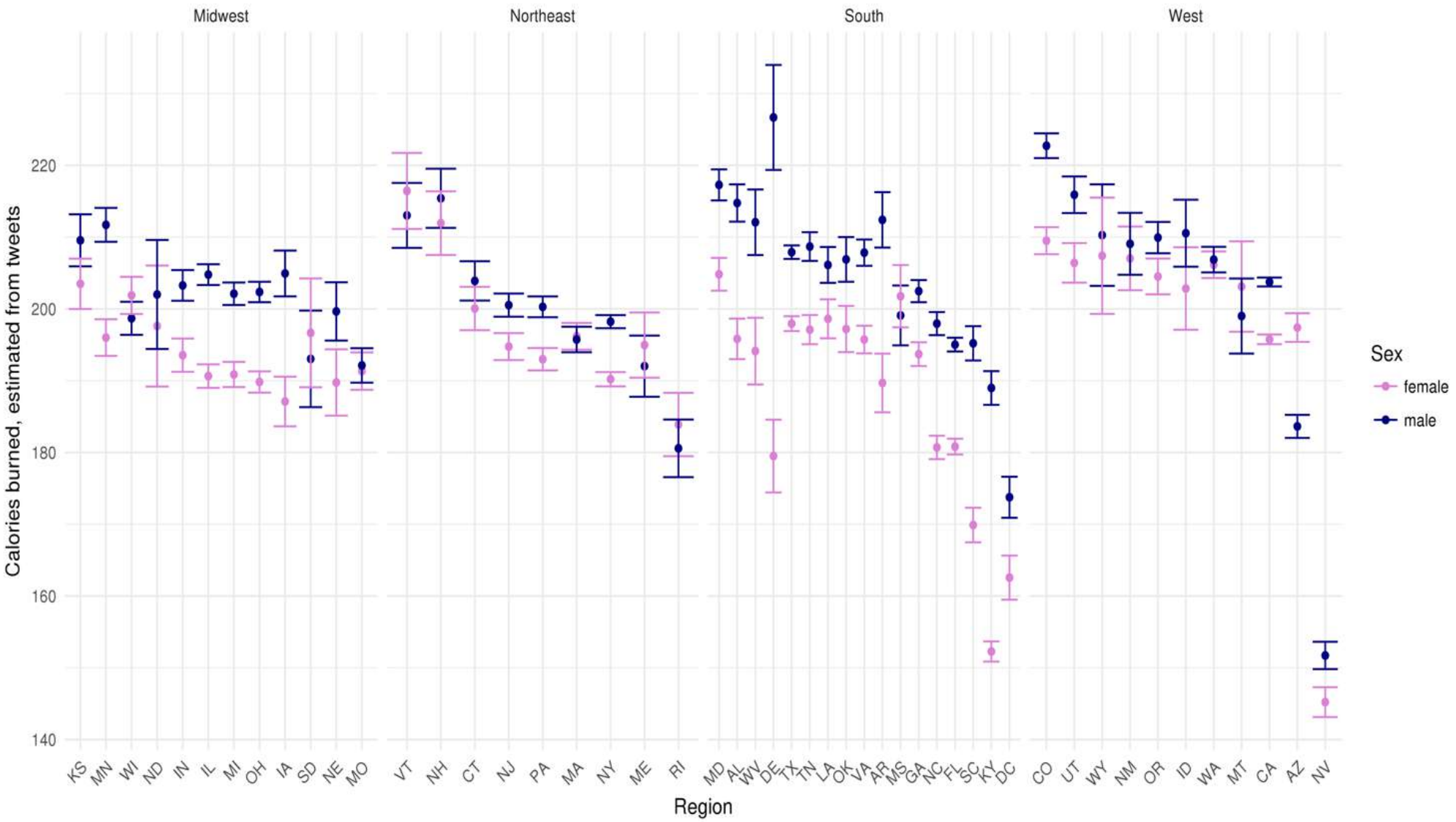
geotagged physical activity tweets

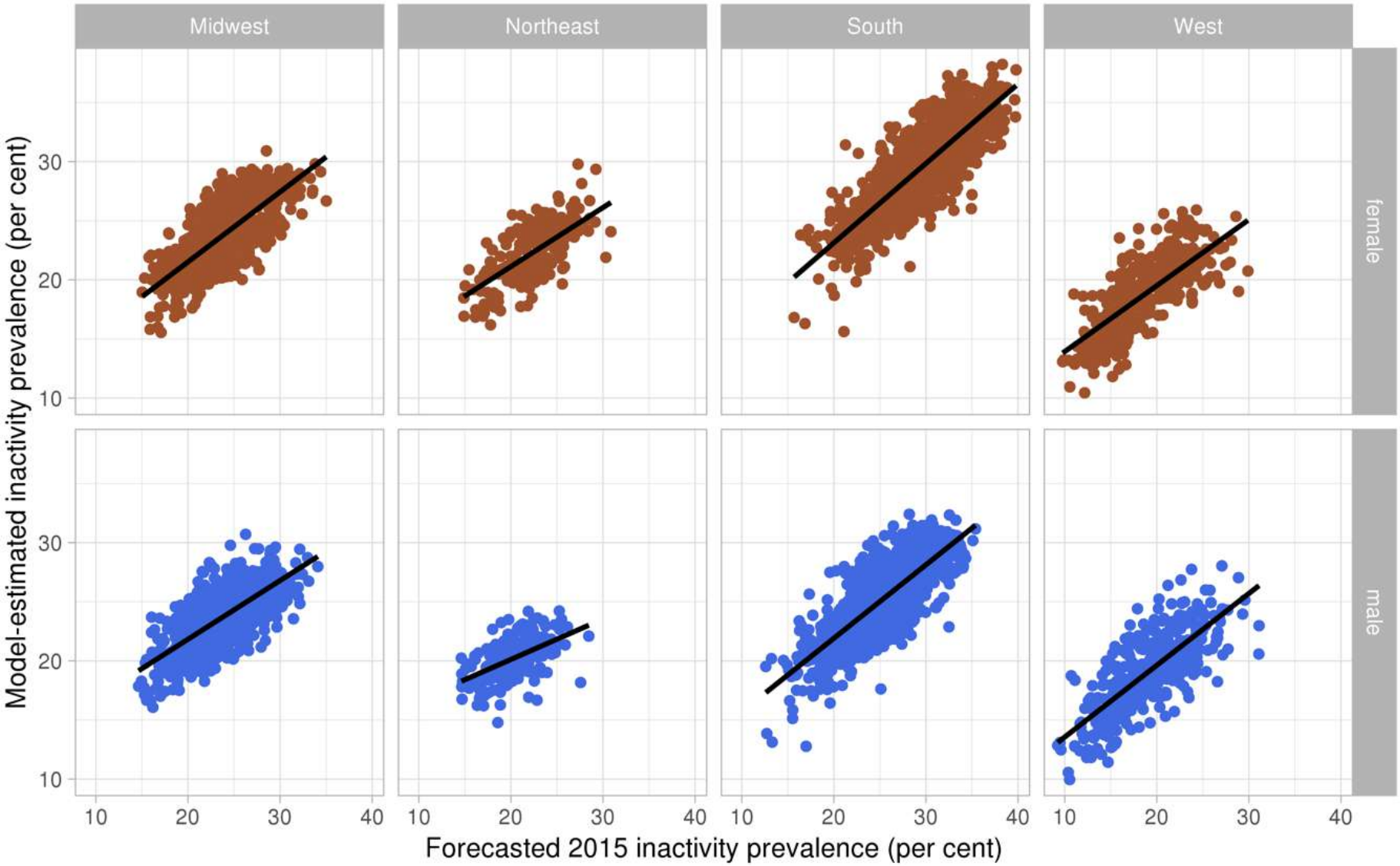
481,146 users

55.7% men and 44.3% women

> 2,900 counties


Most popular activities by sex, region
Midwest
Northeast
South
West
female
male
Per cent of tweets in which activity appears
0.20
0.15
0.10
0.05
0.00
bowl
crossfit
dance
golf
hike
pool
run
swim
walk
workout
yoga
Activity
Activity
bowl
crossfit
dance
golf
hike
pool
run
swim
walk
workout
yoga


Midwest
Northeast
South
West
Calories burned, estimated from tweets
220
200
180
160
140
KS MN WI ND IN IL MI OH IA SD NE MO
VT NH CT NJ PA MA NY ME RI
MD AL WV DE TX TN LA OK VA AR MS GA NC FL SC KY DC
CO UT WY NM OR ID WA MT CA AZ NV
Region
Sex
female
male


Midwest
Northeast
South
West
female
male
Model-estimated inactivity prevalence (per cent)
Forecasted 2015 inactivity prevalence (per cent)
10
20
30
40

Women’s health

Miscarriage

the spontaneous abortion of a viable foetus during the first 20 weeks of pregnancy

Miscarriage facts and misinformation

Although an estimated 15 to 20% of known pregnancies end in a miscarriage, data suggest miscarriages are largely misunderstood and those affected can feel isolated.

Miscarriage facts and misinformation

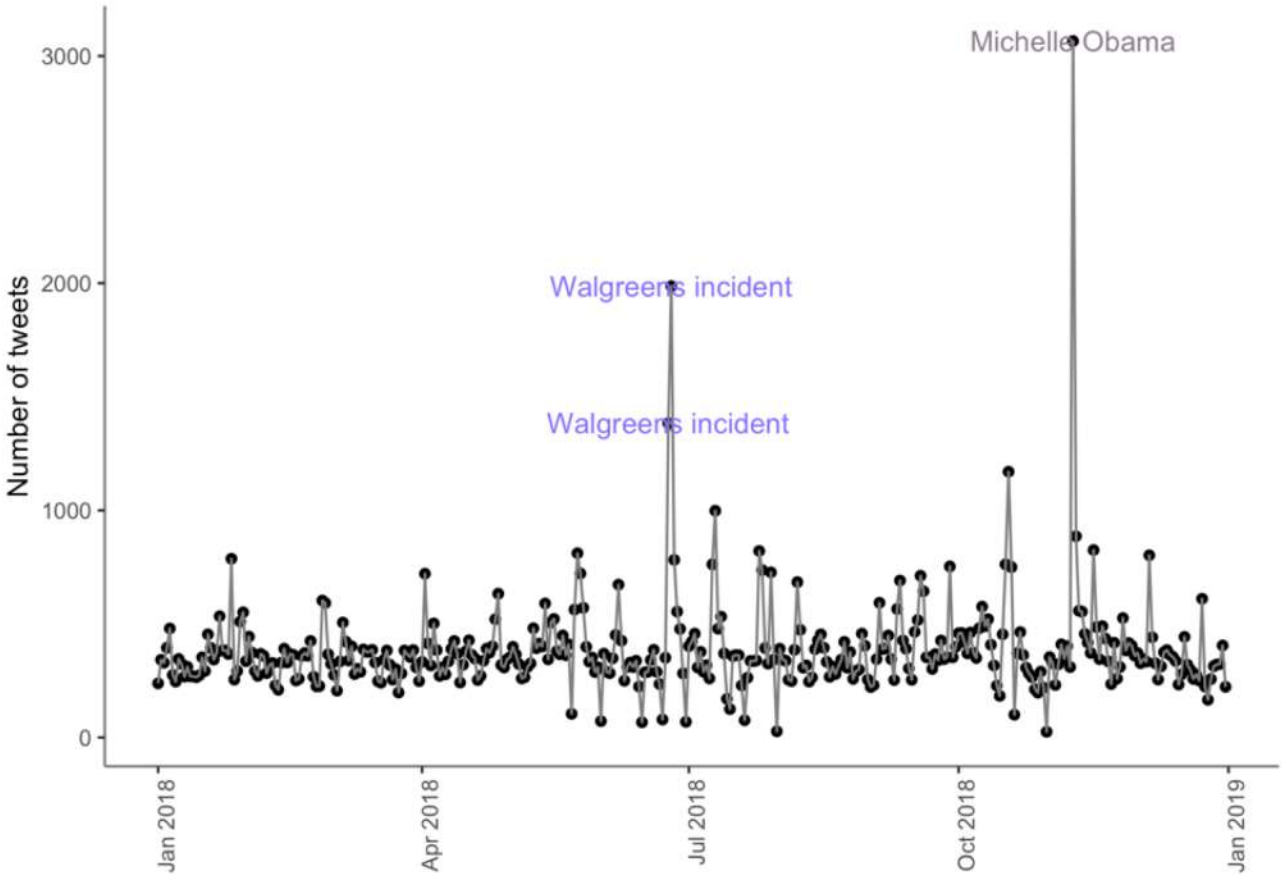
Most miscarriages are due to genetic problems that have nothing to do with witchcraft, environment or behavioral choices.

Data & Methods

- Data from Twitter (291,443 postings)
- Applied Latent Dirichlet Allocation (LDA) to identify major topics of discussion
- Used Locally Weighted Least Squares Regression (LOESS) to identify trends and drivers of discussion
- Manually classified sentiment

Topics of discussion

- Michelle Obama (8.4% of tweets)
- Celebrity (23.0%)
- Preterm birth (10.9%)
- Politics (17.6%)
- Loss and anxiety (10.1%)
- Ectopic pregnancy (7.50%)
- Healthcare (10.7%)
- Influenza vaccine (11.7%)



Discussions of personal experiences of miscarriage by celebrities such as, Michelle Obama can drive conversation on this topic

References

- Cesare N, Oladeji O, Ferryman K, Wijaya D, Hendricks-Muñoz KD, Ward A, **Nsoesie EO**. Discussions of Miscarriage and Preterm Births on Twitter. 2019; (In Press)
- Cesare N, Nguyen QC, Grant C, **Nsoesie EO**. Social media captures demographic and regional physical activity. 2019; *BMJ Open Sport & Exercise Medicine* 2019; 5:e000567. doi:10.1136/ bmjsem-2019-000567.
- Maharana A, **Nsoesie EO**. Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity. *JAMA Network Open*. 2018;1(4):e181535. doi:10.1001/jamanetworkopen.2018.1535.
- Cesare N, Grant C, Hawkins JB, Brownstein JS, **Nsoesie EO**. Demographics in Social Media Data for Public Health Research: Does it matter? *Bloomberg Data for Good Exchange*. 2017 https://arxiv.org/abs/1710.11048
- **Nsoesie EO****, Kraemer MUG** 1, Golding N, Pigott DM, Brady OJ, Moyes CL, Johansson MA, Gething PW, Velayudhan R, Khan K, Hay SI, Brownstein JS. Global Distribution and Environmental Suitability for Chikungunya Virus, 1952 to 2015. *Eurosurveillance*. 2016; 21(20):pii=30234. doi http://dx.doi.org/10.2807/1560- 7917.ES.2016.21.20.30234.

What methods would you use to solve the previously identified health problem? (10 mins)

Ethics

Nextgov

Can Social Media Predict When You'll Die?

FREEEDA/SHUTTERSTOCK.COM

By BOSTON UNIVERSITY // Futurity // OCTOBER 1, 2019

Could social media data predict your death?

SOCIAL MEDIA

Social media platforms like Twitter, Instagram, and

Should your social media data be used to predict when you'll die?

Just because you can explore a research topic/question, doesn't mean you should.

Have your study participants consented to how you collect and use their data?

Who is your data?

Who benefits from your study?

10.28.19

Technology biased against black patients runs rampant in hospitals

A new study shows that a widely used algorithm for predicting which patients get additional care is disproportionately counting out black patients—and could have left tens of thousands without adequate medical care.

Is your data biased towards one group?

Representation does not equal equity.

Exercise

What are the next steps for your project?

Contributors

Too many ☺

Thank you!

Contact
Email: onelaine@bu.edu
Twitter: @ensoesie